Exp. Test

1
mc

Research Memorandum 65-6

CONSTRUCTION AND EVALUATION OF TWO ABBREVIATED ENLISTMENT SCREENING TESTS

August 1965

APRO

U. S. ARMY PERSONNEL RESEARCH OFFICE

(12) 12

Army Project Number
2J024701A722

Input Quality c-16

(9) Research Memo[redacted] 65-6

(6)

CONSTRUCTION AND EVALUATION OF TWO ABBREVIATED ENLISTMENT SCREENING TESTS.

(10) ~~by~~ Leonard C. Seeley

A. G. Bayroff, Task Leader

(14) APRO-RM-65-6

Submitted by:
Edmund F. Fuchs
Chief, Military Selection
Research Laboratory

Approved by:
J. E. Uhlaner
Director of
Laboratories

(11) Aug[redacted]65

Research Memorandums are informal reports on technical research problems. Limited distribution is made, primarily to personnel engaged in research for the U. S. Army Personnel Research Office.

040650 Gm

OBJECTIVE

Since 1953, applicants for enlistment in the Army have been required to pass the Enlistment Screening Test (EST) before being sent to an Armed Forces Examining Station to take the Armed Forces Qualification Test (AFQT). The current forms 3 and 4 of the EST, consisting of 48 items and requiring 45 minutes to administer, have been operational since 1956. These forms yield a full range of percentile scores. However, the requirement is merely to predict pass-fail, that is, to provide "go-no go" information.

The feasibility of drastically shortening the EST has been suggested by research from two quarters. Cronbach and Warrington (3) showed that short multiple-choice tests built of items of uniform difficulty for use with a particular qualifying score can be as valid as longer tests covering a full range of item difficulty. U. S. APRO experience with the Army Qualification Battery (1,4) showed that where concern focused on a narrow range of ability (here, primarily AFQT Category IV), tests could be shortened by omitting items substantially above and below the difficulty level of concern without appreciably impairing effectiveness.

Waters and Heermann (5) applied a theoretical model based on this approach to the problem of developing a shorter EST. Their study involved selection of a small number of items, all at a given level of difficulty--the cut point of concern on the AFQT--rather than a wide range of difficulty as in the current EST. Item responses were correlated with a pass-fail criterion at selected cut points on AFQT (using phi coefficients and tetrachoric r's), and the results were compared with results of similar correlation for the 48-item test. Short tests of 8 items of a given p-value were found to be substantially as effective in predicting pass-fail on AFQT as was the 48-item EST. It was also shown that the a priori method of item selection for the 8-item tests produced as good or higher validity than did more complex test selection procedures. The a priori method capitalized on the known greater correlation with AFQT for the Verbal and Arithmetic Reasoning subtests as against the Pattern Analysis and Shop Mechanics subtests; hence, the short test merely called for four vocabulary items and four arithmetic reasoning items, all with corrected p-values reflecting the cutting score.

The 8-item tests studied by Waters and Heermann were synthetic tests based on item statistics previously obtained when the 8 items were included in 100-item tests. No 8-item test had been administered at any time. The next logical step in the research process was to construct and administer an 8-item test built according to the model and compare the results with those obtained from administration of the 48-item test. Accomplishment of this step was the main purpose of the present study.

CONSTRUCTION OF TWO SHORT EST'S

Two 8-item tests, S-1 (PT 4484) and S-2 (PT 4486) were constructed of items having p-values (corrected for chance success) as near to .69 as possible, in order to reflect the 31st percentile cutting score. In the Waters and Heermann study, a phi coefficient and a tetrachoric r were obtained for each item against a pass-fail criterion on AFQT, and item selection gave consideration to both p-values and validity coefficients against operational AFQT. Since such coefficients were not available for the present study, item selection was based primarily on p-value with secondary consideration given to the item-test biserial correlation coefficient. The items selected for the two short tests, together with sources and available statistics, are shown in Table 1.

VALIDATION OF EST S-1 and S-2

SAMPLES TESTED

The decision was made to base the major portion of the study on selective service registrants rather than on applicants for enlistment --even though the new tests were intended for eventual use with applicants --since applicants would already have taken a form of EST operationally. Use of registrants was judged to be less biasing on the short EST scores than the experience or practice applicants would have gained by taking the longer test.

The two 8-item tests were administered at Armed Forces Examining Stations selected to provide a reasonable geographic coverage--Newark, Baltimore, Columbia, Chicago, Los Angeles. At each AFES, five samples were tested, four of selective service registrants and one of applicants for enlistment.

The total number of cases from the five stations was as follows:

Sample 1: 250 registrants (preinductees without prior military service) given Form S-1

Sample 2: 250 registrants given Form S-2

Sample 3: 200 registrants given 48-item operational EST-4

Sample 4A: 200 registrants given S-1 followed by S-2
4B: 200 registrants given S-2 followed by S-1

Sample 5: 200 applicants for enlistment given Form S-1

The principal purpose of Sample 4 was to determine the equivalence of the two short tests. At the same time, by combining the scores on the 8-item tests, the characteristics of a 16-item test could be estimated.

Table 1

SOURCES AND ITEM STATISTICS FOR EST, S-1 and S-2

Items	Source Test & Item No.		r_{it}	Corrected p-value	S-1 or S-2 Corrected p-value
S-1 Verbal					
1	AFQT 5-6Y	40	.85	.69	.77
2	AFQT 5-6Y	71	.85	.70	.70
3	AFQT 7-8X	101	.85	.68	.60
4	AFQT 8A	35	.82	.69	.74
			Mean 1-4	.69	.70
S-1 AR					
5	AFQT 5-6Z	78	.74	.70	.64
6	AFQT 7-8X	73	.80	.68	.58
7	AFQT 4	65	.79	.70	.67
8	AFQT 6	56	.92	.68	.48
			Mean 5-8	.69	.59
			Mean 1-8	.69	.65
S-2 Verbal					
1	AFQT 6	52	.85	.70	.69
2	AFQT 7-8X	39	.84	.69	.59
3	AFQT 5-6Z	97	.86	.69	.76
4	AFQT 7-8X	71	.83	.68	.66
			Mean 1-4	.69	.68
S-2 AR					
5	AFQT 4	52	.81	.71	.59
6	AFQT 5	56	.85	.69	.58
7	AFQT 5-6Z	105	.80	.70	.60
8	AFQT 7A	40	.85	.67	.74
			Mean 5-8	.69	.63
			Mean 1-8	.69	.65

In all instances, the EST was administered before the operational AFQT. Samples were to be stratified on AFQT (25 cases per decile for samples 1 and 2, 20 per decile for the other samples); hence, considerable over-testing was necessary. AFQT percentile scores were recorded along with Verbal and Arithmetic Reasoning standard scores obtained from the AFQT items to permit computing General Technical (GT) Aptitude Area scores.

Testing was terminated before quotas for Samples 3, 4, and 5 had been filled. The following adjustments were made to complete the samples:

1. In Sample 3, 8 cases were randomly duplicated in the 10th decile, 2 in the 8th, and 4 in the 5th.
2. In Sample 4, order A, 10 cases were randomly duplicated; in order B, 13 cases were randomly duplicated.
3. In Sample 5, no cases were duplicated, but 5 excess cases from the 9th decile were retained against an equal shortage in the 10th, and 13 cases from the 2d decile were retained against an equal shortage in the 1st. These adjustments distorted the distribution in Sample 5, but were at levels where the distortion in phi coefficient, tetrachoric r, and percentage of correct placements should be insignificant.

RESULTS

RELATIONSHIP OF SHORT EST TO AFQT AND GT

AFQT, GT, and EST scores were intercorrelated in samples 1, 2, and 3. The two short EST forms showed very similar results, with correlation coefficients of .80 and .81 against AFQT and .88 and .89 against GT (Table 2). The current operational EST-4 (48 items) produced slightly higher correlation with AFQT (r = .86) but lower correlation with GT (r = .81). In an earlier study by Bayroff and Thomas (2), the 48-item EST produced a coefficient of .83 against AFQT 5 and 6.

In the applicant sample of 200 (Sample 5) administered the S-1 form, coefficients were slightly lower than for registrants--.74 between EST and AFQT, .82 between EST and GT. The more restricted range in this sample as indicated by the smaller standard deviations of all three variables would seem to explain this drop. The restricted range in turn could be a result of the method of adjusting for the shortages in the top and bottom deciles.

Correlation coefficients between the two alternate forms of the EST obtained in Sample 4 were .80 and .84 (Table 3).

Table 2

PREDICTION OF AFQT SCORE
MEANS, STANDARD DEVIATIONS AND PRODUCT MOMENT
CORRELATION COEFFICIENTS FOR FOUR SAMPLES

	M	S.D.	AFQT	GT
Sample 1 (N=250 Registrants)				
AFQT	50.06	29.23		
GT	98.80	22.26	.90	
S-1	5.51	2.53	.81	.89
Sample 2 (N=250 Registrants)				
AFQT	50.21	29.25		
GT	99.02	22.11	.90	
S-2	5.47	2.57	.80	.88
Sample 3 (N=200 Registrants)				
AFQT	49.81	29.14		
GT	99.15	21.86	.89	
EST-4	24.20	11.81	.86	.81
Sample 5 (N=200 Applicants)				
AFQT	50.56	27.09		
GT	101.61	18.85	.87	
S-1	5.91	2.22	.74	.82

RELATIONSHIP OF 16-ITEM TEST TO AFQT AND GT

As stated in the description of the samples, provision was made for obtaining data on a 16-item test. In sample 4, both forms of the short EST were administered to the same individuals; half the sample (4A) were given S-1 followed by S-2; the other half (4B), S-2 followed by S-1. Each individual's scores on the two forms were combined and the resulting 16-item test was correlated with AFQT and GT.

Correlation coefficients of the 16-item test with AFQT were 3 to 6 points higher than those of the 8-item tests. Similarly, the 16-item test yielded coefficients against GT 3 to 6 points higher than those of the 8-item tests. Data are shown in Table 3.

Table 3

INTERCORRELATIONS OF PREDICTOR AND CRITERION SCORES
IN TWO SAMPLES OF 200 REGISTRANTS EACH

	M	S.D.	AFQT	GT	S-1	S-2
Sample 4A (S-1 before S-2)						
AFQT	49.92	29.03				
GT	100.24	22.12	.87			
S-1	5.93	2.34	.75	.84		
S-2	5.94	2.20	.72	.82	.80	
S1+S2	11.87	4.31	.78	.88	.95	.94
Sample 4B (S-2 before S-1)						
AFQT	50.21	29.24				
GT	100.23	22.23	.89			
S-1	5.75	2.52	.80	.86		
S-2	5.55	2.51	.80	.87	.84	
S1+S2	11.30	4.82	.84	.90	.96	.95

PREDICTION OF QUALIFYING SCORE ON AFQT

Major concern of this experiment attached to the ability of the short EST to predict passing or failing the AFQT at the established qualifying score of the 31st percentile. Three statistics were used to indicate this prediction: phi coefficient, tetrachoric r, and percentage of individuals classified the same (pass or fail) by both the EST and the AFQT. Using each of several cutting points on the short EST, the three indices were computed against pass-fail on AFQT for four of the samples. In Sample 3, only two cutting points were used, 25 and the operational qualifying score of 28. Results are shown in Table 4.

Highest validity for pass-fail on the 8-item tests as determined by phi coefficients was for a raw qualifying score of 5 of S-1 and 4 on S-2. This finding was based on the two largest samples (1 and 2) in which the phi's were .75 and .72 respectively. These qualifying scores were supported by tetrachoric r's of .93 and .92, respectively. In Sample 4, the same qualifying scores (5 on S-1, 4 on S-2) resulted when the test of concern was given first. Support for these cutting scores was also found in the percent correctly placed--89 percent in Sample 1 (S-1) and 88 percent in Sample 2 (S-2).

Both the short tests at optimal cutting points were superior to the operational EST-4 at the operational cutting score when evaluated by means of the two coefficients used here as well as by percent correctly placed. With qualifying scores of 5 on S-1 and 4 on S-2, the superiority extended from 8 to 11 correlation points for the phi coefficients, 5 to 6 points for the tetrachoric r's, and 6 to 7 percentage points in terms of correct placements. However, a cutting score of 25 on EST-4 was superior to 28 in this sample, and also as good as results with the short tests.

The 16-item test was somewhat superior to the 8-item tests as indicated by phi coefficients, tetrachoric r's, and percent correctly placed. The highest validity in terms of these three indices was for a qualifying score of 10 when S-1 was given first, or 9 when S-2 was given first. In either case, results were the same: a phi coefficient of .77, an r_{tet} of .95, and 90 percent correctly placed. These values are slightly higher than the corresponding values for the 8-item tests in Sample 4, as well as in Samples 1 and 2.

Table 4

PREDICTION OF AFQT QUALIFYING SCORE
THREE INDICES OF EST PASS-FAIL AGAINST AFQT PASS-FAIL FOR FIVE SAMPLES

Passing Score	phi	r-tet	%placed correctly	Passing Score	phi	r-tet	%placed correctly		Passing Score	phi	r-tet	%placed correctly
Sample 1 (S-1)				Sample 4[a] Order A					Sample 4 Order B			
3	.60	.87	84	3	.56	.90	82	(S-1)	3	.60	.92	84
4	.73	.93	89	4	.68	.93	87		4	.68	.90	87
5	.75	.93	89	5	.74	.94	90		5	.74	.92	89
6	.68	.90	84	6	.66	.87	86		6	.74	.93	88
7	.58	.88	76	7	.55	.80	77		7	.64	.89	82
Sample 2 (S-2)				3	.50	.87	80	(S-2)	3	.58	.87	83
3	.65	.91	86	4	.53	.80	82		4	.73	.93	89
4	.72	.92	88	5	.73	.92	89		5	.71	.91	88
5	.70	.90	87	6	.70	.90	87		6	.69	.92	84
6	.71	.92	86	7	.56	.82	77		7	-	-	-
				6[b]	-	-	-	(S-1+S-2)	6	.60	.88	84
Sample 3 (EST-4)				7[b]	-	-	-	(16 items)	7	.72	.93	88
25	.78	.95	90	8	.65	.90	86		8	.76	.94	90
28	.64	.87	82	9	.71	.92	88		9	.77	.95	90
Sample 5 (S-1)				10	.77	.95	90		10	.76	.94	90
3	.47	.82	79	11	.72	.92	88		11	.73	.92	88
4	.58	.87	83	12	.63	.84	84		12	.74	.94	88
5	.62	.85	84	13	.49	.72	78		13[b]	-	-	-
6	.58	.81	80									
7	.52	.79	74									

[a] In Sample 4, Order A=S-1 before S-2; Order B=S-2 before S-1.
[b] Zero in one of 4 cells.

THE P-VALUES

As previously stated, the items comprising the short EST's were selected to provide p-values as close to .69 as possible. Inasmuch as these p-values had been based on several different samples obtained from five to twelve years ago when the items were part of 100- or 300-item tests, the values were recomputed on the present samples. They are shown in Table 1. For both S-1 and S-2, the mean p-value of the 8 items was approximately .65, just four points below the .69 originally obtained. When items were examined by type, however, the Verbal items maintained on the average the original p-value, whereas the Arithmetic Reasoning items fell to .59 in one form and .63 in the other. Why this should occur is not clear; one possibility to be investigated is that the time limit allowed (5 minutes) was insufficient, and haste on the part of some examinees resulted in increased errors on the items placed last, which were Arithmetic Reasoning items.

CONCLUSIONS

The 8-item EST, when correlated with scores on AFQT, yielded product moment correlation coefficients only slightly lower than that obtained with the 48-item EST-4. For predicting the 31st percentile qualifying score on AFQT, the best cutting scores on the 8-item tests were superior to the operational cutting score on the 48-item test as shown by three indices: phi coefficient, tetrachoric r, and percentage of men placed correctly. The validity of the 16-item test was slightly higher than that of the 8-item test on all indices. It would seem, therefore, that short tests made up of items of nearly the same p-value and appropriate to the qualification score of AFQT might be a suitable replacement for the operational 48-item EST.

Since the guessing factor is likely to have serious impact on such a short test, it was decided to edit these short EST forms to reduce the guessing factor. All items were changed from 4-choice to 5-choice. Another tryout, with less elaborate sampling, will be given these edited forms.

REFERENCES

Bayroff, A. G. and Seeley, L. C. Development of the Army Qualification Battery, AQB-1. U. S. APRO Technical Research Report 1117. October 1959.

Bayroff, A. G. and Thomas, J. A. Evaluation of EST for predicting AFQT performance. U. S. APRO Technical Research Report 1114. February 1959.

Cronbach, L. J. and Warrington, W. G. Efficiency of multiple choice tests as a function of spread of item difficulties. Psychometrika, 17, 1952. pp. 127-147.

Seeley, L. C. and Anderson, A. A. Development of the Army Qualificaticn Battery, Forms 2 and 3. U. S. APRO Technical Research Report 1131. May 1963

Waters, C. J. and Heermann E. F. Feasibility of abbreviated forms for the Enlistment Screening Test. U. S. APRO Technical Research Note 144. May 1964.